Recent Developments and Future Directions in Bayesian Model Averaging

David Kaplan
Department of Educational Psychology

THE UNIVERSITY of WISCONSIN MADISON

Indiana University Workshop in Methods Series
February 7th 2020

Introduction

- Outline of Talk
  - A gentle (and brief!) introduction to Bayes
  - The Bayesian advantage
  - Brief history of BMA research
  - Method of BMA
  - Extensions of BMA
  - Future directions
  - Conclusions

- Notation
  - $p(y|\theta)$: **Distribution of the current data given the model parameters**. Parameters are unknown but assumed fixed. This would be your statistical model of interest. Frequentist (classical statistics) stops here.
  - $p(\theta)$: **Prior probability distribution of the parameters**. Parameters are unknown, but assumed to be random variables. This is where your epistemic uncertainty is encoded.
  - $p(\theta|y)$: **Posterior distribution of the parameters.** This is where you summarize your new knowledge about model parameters described in terms of probability distributions.

Bayes' Theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1}$$

- The prior distribution, reflecting uncertainty in prior knowledge, is combined with the distribution of the current data given the parameters and the model to yield updated estimates of parameters through Bayes' Theorem.

The Bayesian Advantage

- The major advantages of Bayesian statistical inference over frequentist statistical inference are
  - **Inferences are based on data actually observed.**
  - **Focuses on quantifying evidence.**
  - **Supports evolutionary knowledge building through Bayesian updating.**
  - **Allows exploring the full posterior distribution of an effect.**
  - **Offers a reasonable alternative to the NHST.**

Bayesian Model Averaging

- Within the Bayesian framework, parameters are not the only unknown elements.
- The Bayesian framework recognizes that model selection is a decision taken under incomplete information (uncertainty).
- The uncertainty manifests itself in the choice among competing models.
- The uncertainty lies in not knowing (never knowing) the true data generating model.
- This form of uncertainty often goes unnoticed – particularly in the social and behavioral sciences.

Bayesian Model Averaging

- Quoting Hoeting et al. (1999)

> *"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(pg. 382)*

- An internally consistent Bayesian framework for modeling and estimation must also account for model uncertainty.

- One popular approach to addressing the problem of uncertainty lies in the method of *Bayesian model averaging* (BMA; Leamer, 1978; Madigan & Raftery, 1994; Raftery, 1997; Hoeting et al. 1999; Clyde,1999, Draper, 1995).

Method of BMA

- Let Υ, represent a quantity of interest such as a predicted value.

- Next, consider a set of competing models $\mathcal{M}_k$, $k = 1, 2, \ldots, K$ that are not necessarily nested.

- The posterior (mixture) distribution of Υ given data y can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|\mathcal{M}_k)p(\mathcal{M}_k|y). \tag{2}$$

where $p(\mathcal{M}_k|y)$ is the posterior probability of model M_k written as

$$p(\mathcal{M}_k|y) = \frac{p(y|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{l=1}^{K} p(y|\mathcal{M}_l)p(\mathcal{M}_l)}, \qquad l \neq k. \tag{3}$$

- $p(\mathcal{M}_k|y)$ will likely be different for different models.

Method of BMA

- The term $p(y|\mathcal{M}_k)$ can be expressed as an integrated likelihood

$$p(y|\mathcal{M}_k) = \int p(y|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k, \tag{4}$$

  where $p(\theta_k|\mathcal{M}_k)$ is the prior distribution of θ_k under model M_k (Raftery et al., 1997).

- Thus, BMA provides an approach for combining models specified by researchers or combined algorithmically.

- The advantage of BMA has been discussed in Madigan and Raftery (1994), who showed that BMA provides better out-of-sample (and long-run frequency) predictive performance than that of any single model based on rules that describe predictive success.

Method of BMA

- Practical concerns:
  1. The number of terms in $p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|\mathcal{M}_k)p(\mathcal{M}_k|y)$ can be quite large and the corresponding integrals are hard to compute.
  2. Eliciting $p(\mathcal{M}_k)$ may not be straightforward. The uniform prior $1/\mathcal{M}$ is often used, but can be modified.
  3. Choosing the class of models to average over is also challenging.
- The problem of reducing the overall number of models has led to solutions based on *Occam's window* or the *Markov chain Monte Carlo Model Composition* (MC^3) algorithm, among others. (Madigan and Raftery, 1994).

Method of BMA

- The default prior on the model parameters is the *unit information prior* (Kass & Raftery, 1995; Raftery, 1998).
  - A special case of Zellner's g-prior.
  - A weakly informative (data-based) prior that is diffused over the region of the likelihood where parameter values are considered mostly plausible, but not overly spread out.
- Priors on the model space are equivalent for all models: $1/\mathcal{M}$.

Validating BMA: Scoring Rules

- A key characteristic of statistics is to develop accurate predictive models (Dawid, 1984).
- All other things being equal, a given model is to be preferred over other competing models if it provides better predictions of what actually occurred. **Again, think weather forecasting.**
- We need to decide on rules for gauging predictive accuracy – *scoring rules*.
- Scoring rules provide a measure of the accuracy of probabilistic forecasts.
- A forecast can be said to be "well-calibrated" if the assigned probabilities of the outcome match the actual proportion of times that the outcome occurred. Note the frequentist interpretation.

Three examples of scoring rules

- For continuous distributions, we can use the Kullback-Liebler divergence.

$$I(f,g) = \int f(x) log \left(\frac{f(x)}{g(x|\theta)} \right) dx \quad (5)$$

- A related measure to KL-divergence is the log predictive score.

$$-\sum_i \log \left[p(y_i^f | X, y, X_i^f) \right] \quad (6)$$

- For a dichotomous outcome we can use the Brier score defined as

$$Brier = \frac{1}{T} \sum_{t=1}^{T} (f_t - o_t)^2, \quad (7)$$

where f_t is the probabilistic forecast and o_t is the observed event (1 if the forecasted event took place, 0 otherwise).

An Important Assumption: $\mathcal{M}$-frameworks

- Clyde & Iversen (2012), Vehtari & Ojanen (2012)
- We have assumed that the true data generating model, say, $\mathcal{M}_T$ is one of the models in the set of models $\mathcal{M} = \{\mathcal{M}_k, k = 1, 2, \ldots, K\}$.
- This assumption is referred to as the $\mathcal{M}-closed$ framework.
- In the $\mathcal{M}-closed$ framework, it makes sense to assign prior probabilities that $\mathcal{M}_T$ is in the space of models.
- $\mathcal{M}_T$ is usually not available to us which is why we have model uncertainty.
- BMA will not get us the true DGM, but only the one closest to the true model in terms of KL-divergence.

An Important Assumption: $\mathcal{M}$-frameworks

- In the $\mathcal{M}-complete$ framework, the analyst has entertained a true DGM M_T but that this true model lies outside the set of models $\mathcal{M} = \{\mathcal{M}_k, k = 1, 2, \ldots, K\}$being considered.
- We still consider the set of models in $\mathcal{M}$ because they are easier to communicate
- In the $\mathcal{M}-open$ frameworks $\mathcal{M}_T$ is not in the set of models $\mathcal{M}$, and we either lack the ability, time, or expertise to specify it.
- These frameworks are important with respect to the assignment of prior model probabilities.

Example: BMA in regression analysis

See file: BMA.html

BMA for Propensity Score Analysis

Kaplan & Chen (2012), *Psychometrika*; Kaplan & Chen (2014), *MBR*; Chen & Kaplan (2015), *JREE*

- In observational studies, individuals self-select into treatment conditions on the basis of an unknown mechanism.
- The selection process is often highly nonrandom, introducing selection bias that may result in highly unbalanced covariates and thus severely weakening causal inferences.
- Establishing balance between the treatment and control groups on observable covariates is thus essential for obtaining unbiased treatment effect estimates.

BMA for PSA

- Propensity score analysis is a well-established method for obtaining balanced treatment and control groups.

- Let z be a set of covariates that predict selection into treatment. The estimated propensity score can be written as

$$\hat{e}(z) = p(T = 1|z), \tag{8}$$

- The estimated propensity score $\hat{e}(z)$ has many important properties. Perhaps the most important property is the *balancing* property, which states that those in $T = 1$ and $T = 0$ with the same $\hat{e}(z)$ will have the same distribution on the covariates z.

- Formally, the balancing property can be expressed as

$$p\{z|T = 1, \hat{e}(z)\} = p\{z|T = 0, \hat{e}(z)\}, \tag{9}$$

BMA for PSA

- Kaplan & Chen (2012) proposed a two-step Bayesian propensity score approach that:
  1. Separates the PS equation from the outcome equation.
  2. Allows priors to be incorporated into the PS and outcome equations.
  3. Shows excellent covariate balance (Chen and Kaplan, 2015).
  4. Shows excellent frequentist properties (*calibrated Bayes*; Little, 2006; 2011; 2012).

BMA for PSA

- A problem is that our Bayesian propensity score approach assumes that the propensity score model itself is fixed.
- Rather, as a model for treatment selection, it is reasonable to assume that many possible models could have been chosen.
- Moreover, the goal in Bayesian model selection is to choose a model that has the best predictive capacity. Here we are trying to optimally predict selection into treatment.
- Kaplan & Chen (2014) argued that a full accounting of uncertainty in propensity score analysis should address model uncertainty and optimize prediction, and thus explored Bayesian model averaging in the propensity score context.

Summary of Simulation Study

- Differences across all conditions are modest.
- Occam's window does not influence treatment effect estimation for our Bayesian model averaging approach.
- Priors on the propensity score model parameters have little impact on the treatment effect estimation.
- BMA works best for stratification, optimal matching, and regression.
- BMA produces better estimates of treatment effects than the two-step BPSA which does not account for model uncertainty.

BMA for SEM

Kaplan & Lee (2015), *SEM*; Kaplan & Lee (2018), *Evaluation Review*

- The general steps of our BMA-SEM method are as follows

  1. Specify an initial model of interest recognizing that this may not be the model that generated the data.

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \zeta \tag{10}$$

  2. Reduce the model space using an MC^3-type algorithm, treating a path diagram as a DAG.

  3. Obtain the weighted average of structural parameters over each model, weighted by the PMPs by transforming the structural model to a regression model.

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1} + (\mathbf{I} - \mathbf{B})^{-1}\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\zeta \tag{11}$$

Results of Simulation Study

- No difference among methods when the true model is known. Scoring rules give same results.
- No influence of width of Occam's window on predictive coverage.
- Our method performs no worse than when the true model is known.
- Our simulation study establishes the groundwork for applying the method to real data.

Results of Case Study

- For the case study, we use data from PISA 2009 to estimate a model relating reading proficiency to a set of background and reading strategy variables.
- The sample was collected from PISA-eligible students in the United States, and the sample size was 5,053.
- The sample was split into a model averaging set (N = 2,526) and a predictive testing set (N = 2,527).

Results of Case Study

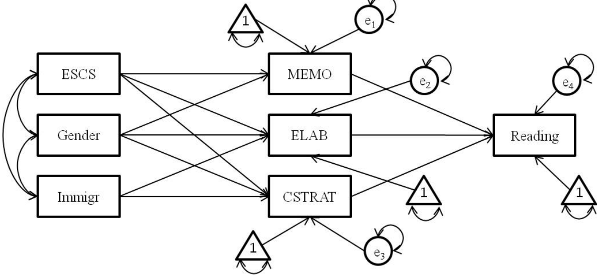


Figure 1: Initial path model.

Table 1: Selected models by BMA-SEM with the C = 100 for the PISA data

Parameter[a]	M_1	M_2	M_3
MEMO∼ESCS	●	●	
ELAB∼ESCS	●	●	●
CSTRAT∼ESCS	●	●	●
Reading∼ESCS	●	●	●
MEMO∼Gender	●	●	●
ELAB∼Gender	●	●	●
CSTRAT∼Gender	●	●	●
Reading∼Gender	●	●	●
MEMO∼Immigr	●	●	●
ELAB∼Immigr			
CSTRAT∼Immigr		●	
Reading∼Immigr			
ELAB∼MEMO	●	●	●
CSTRAT∼MEMO	●	●	●
Reading∼MEMO	●	●	●
CSTRAT∼ELAB	●	●	●
Reading∼ELAB	●	●	●
Reading∼CSTRAT	●	●	●
BIC	39461.68	39464.74	39465.15
PMP	0.72	0.15	0.13

Note. [a] ∼ refers to regression of left-hand variable onto right-hand variable; PMP = posterior model probability.

Table 2: Ninety percent coverage for PISA example

Method	90% Coverage
BMA-SEM (4)	0.90
BMA-SEM (20)	0.90
BMA-SEM (100)	0.90
FSEM	0.88
BSEM	0.88

Recent Extensions: BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

- Multiple imputation (MI) Rubin (1987) is arguably the gold-standard in addressing problems of missing data.
- MI can be implemented under three general approaches
  - Monotone missing data imputation
  - Joint modeling
  - Fully conditional specification (chained equations)

BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

- Overarching the various methods of MI, Rubin (1987) introduced the idea of so-called “proper" imputations.
- Proper imputations relate to the asymptotic properties of the statistics of interest obtained from the imputation process over an infinitely large number of imputed data sets.
- If Rubin’s (1987) combining rules yield a consistent and asymptotically normal estimator, then the imputations are “proper" (Rubin, 1987).

BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

- Under the Bayesian framework, let Y^{mis} represent observations on Y that are missing, and let Y^{obs} represent observations on Y that are observed.

- The posterior predictive distribution of the missing data can be written as

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|Y^{obs}, \theta)p(\theta|Y^{obs})d\theta, \tag{12}$$

- As long as imputations are the result of independent realizations of equation (12), they are said to be “Bayesianly proper" (Shafer, 1997).

BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

- Inferences are “Bayesianly proper" because parameter uncertainty is being addressed in the imputation process through the priors placed on the model parameters.
- However, model parameters are not the only sources of uncertainty in the imputation process.
- There is uncertainty in the choice of the imputation model and this uncertainty is not being accounted for in conventional Bayesianly proper multiple imputation.
- For multiple imputation to be fully Bayesianly proper, it is necessary to account for imputation model uncertainty.

BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

- We apply Bayesian model averaging to multiple imputation under the chained equations approach to Bayesian normal theory-based multiple imputation.
- Our approach simply adds a Bayesian model averaging component to each cycle through the chained equations.
- As each variable takes its turn as the target variable for imputation, Bayesian model averaging is applied to the imputation model for that target variable.
- Imputation model uncertainty (as well as parameter uncertainty) is accounted for across all variables and iterations.
- Our approach is “really", “truly", “Bayesianly" proper.

BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

Table 3: Descriptive Statistics for Posterior Model Probabilities for miBMA

			MAR			MCAR				
			MAR1	MAR2	MAR3	MCAR1	MCAR2	MCAR3	MCAR4	MCAR5
N=100	Cor 0.2	20%	0.581	0.578	0.584	0.572	0.573	0.574	0.572	0.574
		40%	0.614	0.615	0.616	0.609	0.612	0.611	0.609	0.610
		60%	0.639	0.638	0.640	0.637	0.638	0.638	0.638	0.637
	Cor 0.6	20%	0.562	0.564	0.573	0.567	0.568	0.568	0.569	0.571
		40%	0.613	0.614	0.619	0.609	0.609	0.609	0.611	0.610
		60%	0.653	0.654	0.654	0.645	0.647	0.644	0.643	0.640
N=1000	Cor 0.2	20%	0.719	0.725	0.720	0.706	0.705	0.701	0.703	0.699
		40%	0.828	0.829	0.830	0.837	0.838	0.839	0.839	0.838
		60%	0.914	0.914	0.917	0.923	0.925	0.924	0.927	0.925
	Cor 0.6	20%	0.854	0.852	0.855	0.867	0.871	0.865	0.863	0.869
		40%	0.888	0.887	0.888	0.899	0.897	0.891	0.891	0.897
		60%	0.940	0.938	0.939	0.961	0.962	0.961	0.961	0.962
N=5000	Cor 0.2	20%	0.985	0.986	0.987	0.981	0.983	0.982	0.982	0.982
		40%	0.974	0.973	0.974	0.972	0.975	0.972	0.973	0.972
		60%	0.985	0.985	0.984	0.994	0.994	0.993	0.994	0.994
	Cor 0.6	20%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		40%	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.999
		60%	0.997	0.997	0.997	0.992	0.993	0.993	0.993	0.993

Note: Values in the table are averaged over all iterations, imputations, and replications.

BMA for Missing Data

Kaplan & Yavuz, (2019) *MBR*

Table 4: Selected KL and MSE Results for MAR Mechanism

				KL-Test			MSPE		
		Missing	Method	MAR 1	MAR 2	MAR 3	MAR 1	MAR 2	MAR 3
N=100	Cor 0.2	20%	miBMA	0.048	0.050	0.048	2.050	2.054	2.031
			NORM	0.055	0.057	0.055	2.125	2.110	2.114
		40%	miBMA	0.078	0.083	0.082	2.102	2.116	2.091
			NORM	0.087	0.094	0.088	2.335	2.348	2.316
		60%	miBMA	0.106	0.108	0.109	2.256	2.290	2.253
			NORM	0.115	0.120	0.116	2.982	2.975	2.896
	Cor 0.6	20%	miBMA	0.056	0.058	0.058	1.085	1.076	1.068
			NORM	0.060	0.061	0.061	1.099	1.085	1.089
		40%	miBMA	0.066	0.073	0.070	1.109	1.120	1.108
			NORM	0.073	0.077	0.075	1.208	1.217	1.202
		60%	miBMA	0.080	0.084	0.085	1.203	1.215	1.196
			NORM	0.091	0.096	0.095	1.535	1.556	1.510
N=5000	Cor 0.2	20%	miBMA	0.013	0.015	0.016	1.733	1.740	1.734
			NORM	0.013	0.015	0.016	1.732	1.737	1.734
		40%	miBMA	0.020	0.022	0.024	1.744	1.746	1.743
			NORM	0.020	0.022	0.024	1.742	1.743	1.742
		60%	miBMA	0.027	0.030	0.031	1.754	1.757	1.756
			NORM	0.028	0.030	0.031	1.752	1.754	1.754
	Cor 0.6	20%	miBMA	0.022	0.022	0.022	0.894	0.896	0.894
			NORM	0.022	0.022	0.023	0.895	0.897	0.895
		40%	miBMA	0.023	0.023	0.023	0.899	0.900	0.899
			NORM	0.022	0.022	0.022	0.901	0.901	0.901
		60%	miBMA	0.023	0.022	0.023	0.905	0.907	0.906
			NORM	0.023	0.023	0.023	0.907	0.907	0.908

BMA for Forecasting with LSAs

Kaplan & Huang (in progress)

- The National Assessment of Educational Progress (NAEP) has provided long-term trend data since 1970.
- At the state level, NAEP has provided long-term trend data since 1996, and particularly after 2001 with the reauthorization of the Elementary and Secondary Education Act.
- NAEP can provide important monitoring and forecasting information regarding population-level academic performance.
- NAEP can provide student/school demographic and student “non-cognitive" information via survey questionnaires that can be used as outcomes in their own right, and as predictors of the shape of trends in academic outcomes over time.

BMA for Forecasting with LSAs

Kaplan & Huang (in progress)

- At the state/jurisdiction level, NAEP provides panel data that can be used to model trends in educational outcomes over time.

- The purpose of this work is to develop a "proof-of-concept" that the state NAEP assessments can be used to specify cross-state growth regressions and to develop probabilistic predictive models that can be used to forecast trends across states in important educational outcomes.

- To obtain optimal predictive probabilistic models, we will utilized a set of statistical methodologies situated within the Bayesian paradigm of statistics to explicitly address issues of uncertainty in model parameters estimation and model choice.

BMA for Forecasting with LSAs

Kaplan & Huang (in progress)

- This work is motivated by the work of Fernadez, Ley, & Steele (2001) who developed Bayesian model-averaged (BMA) growth regressions for gross domestic product over 140 countries.
- Fernadez, Ley, & Steele found BMA-based growth regressions to be superior to any single model chosen on the basis of out-of-sample-predictive performance.
- Similarly, we propose to estimate a latent growth regression model of NAEP 8^{th} grade mathematics performance across the 50 states and the District of Columbia.
- We provide predictive densities of growth allowing comparison of the actual growth rate in mathematics achievement and the growth rate predicted by the model.
- We will explore these models for 8^{th} boys, girls, white students, and non-white students

Analysis Steps

1. We estimate simple growth regression of 8^{th} grade mathematics achievement without the inclusion of predictors. We estimate these growth regressions separately for boys and girls and Whites and Non-Whites.

2. We select a set of variables deemed to be important predictors of the shape of the trends over time. We will also be guided by an inspection of the student, teacher, and school questionnaires and other data sources.

3. We add the policy relevant predictors to the model. In this step, we calculate difference scores between the 2017 and 2003 measures of these variables (if available) and use these change scores as predictors of growth in 8^{th} mathematics achievement.

4. We implement BMA to our forecasting model and compare the results to the model estimated in Step 3. We first compute the random growth parameters for each state/jurisdiction and import them into the BMA software.

5. Finally, we use the results of the Bayesian model averaging step to obtain posterior predictive densities of growth rates in 8^{th} mathematics achievement for the four groups across the states.

    - These predictive densities provide a means of checking the prediction model of the growth rate against the actual growth. Also, 95% posterior intervals around the predicted growth rates are provide.

Bayesian growth curve modeling

- We use latent growth curve modeling treating US states and DC as the unit of analysis and estimating the trajectories in educational outcomes over time.

- We write the intra-state model as

$$y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + r_{ti} \tag{13}$$

  where y_{it} is the outcome for state i $(i = 1, \ldots n)$ at time t $(t = 1, \ldots T)$ where a is the NAEP year of the assessment, π_{0i} is the intercept capturing state i's status on the outcome at time t, π_{1i} is the slope for state i at time t, and r_{it} is the residual term.

- We allow non-linear “spline" estimation of some time coefficients. The first three time points (2003, 2005, and 2007) are fixed while the remaining time points are estimated from the data.

Bayesian growth curve modeling

- The model in Equation (13) is flexible enough to allow the growth parameters to be predicted by state level time-invariant covariates.

$$\pi_{pi} = \beta_{p0} + \sum_{k=1}^{K_p} \beta_{pk} x_{ki} + \epsilon_{pi}, \tag{14}$$

where the π_{pi} are the growth parameters (intercept and growth rate), x_{ki} are values on K predictors for state i, and ϵ_{pi} are errors.

- Priors are added to all model parameters, and the model can/will be estimated via the “blavaan" interface to “rjags".

Data source

- Two data sources were combined to provide the variables necessary for this analysis.

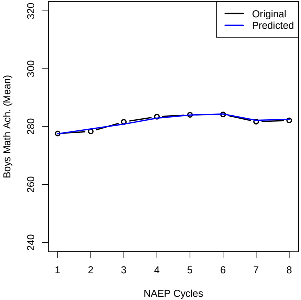
  1. NAEP state mathematics achievement and reading data from 2003 to 2017, National School Lunch Program variables obtained as the percentage of students who are NSLP-eligible, and taken as an SES proxy. Demographic variables such as percentage of gender and race/ethnicity groups were also included in this data file.
  2. The NAEP data file was merged with specific variables in the Common Core of Data (CCD) to obtain information regarding state staff counts (FTEs), per pupil state revenue, pupil/teacher ratio. It should be noted that Tennessee was excluded from this analysis due to lack of state reported data in the CCD file.

Table 5: Bayesian growth curve modeling results for boys & girls

Boys growth params.	Estimate	Post.SD	HPD.025	HPD.975	PSRF	Prior
intercept	274.782	1.300	272.116	277.226	1.000	dnorm(260,.1)
slope	0.931	0.108	0.726	1.149	1.000	dnorm(0,1e-2)
Pre(intercept)	88.460	19.415	55.216	127.007	1.000	dwish(iden,3)
Pre(slope)	0.246	0.067	0.132	0.385	1.000	dwish(iden,3)
Pre(Intercept,slope)	-2.614	0.884	-4.413	-1.079	1.000	dwish(iden,3)

Girls growth params.	Estimate	Post.SD	HPD.025	HPD.975	PSRF	Prior
intercept	273.707	1.308	271.113	276.219	1.000	dnorm(260,.1)
slope	0.902	0.122	0.669	1.142	1.002	dnorm(0,1e-2)
Pre(intercept)	85.114	19.065	51.593	122.738	1.000	dwish(iden,3)
Pre(slope)	0.269	0.084	0.126	0.43	1.000	dwish(iden,3)
Pre(Intercept,slope)	-2.826	0.982	-4.749	-1.113	1.000	dwish(iden,3)

BGCM results



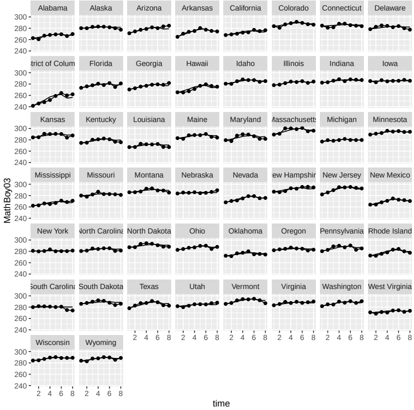


Figure 2: Country-level and state-level fitted trend in 8^{th} grade boys mathematics achievement.

BGCM results

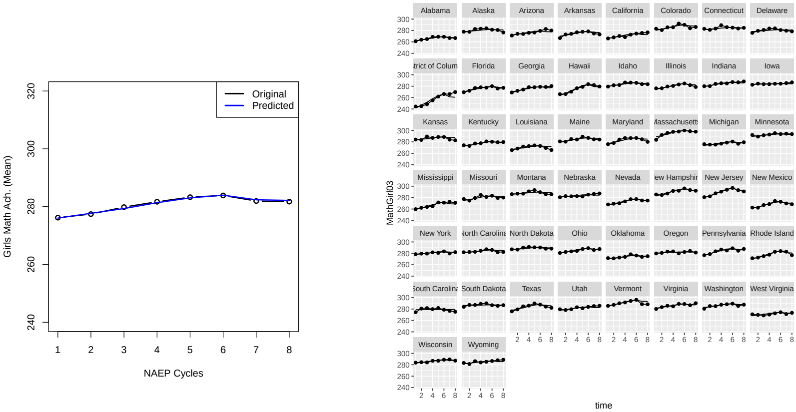


Figure 3: Country-level and state-level fitted trend in 8^{th} grade girls mathematics achievement.

BMA results

Table 6: Bayesian Model Averaging Results for Boys & Girls

Boys	PIP	Post Mean	Post SD	Cond. Pos. Sign
ReadBoyDiff	1.00	0.05	0.01	1.00
TOTREVdiff	0.21	0.05	0.13	1.00
PTRatioDiff	0.19	-0.01	0.03	0.00
BoysEnrollDiff	0.16	0.01	0.02	1.00
NSLPLunchDiff	0.12	0.00	0.00	0.87
FTEdiff2	0.12	-0.00	0.00	0.37

Girls	PIP	Post Mean	Post SD	Cond. Pos. Sign
ReadGirlDiff	0.98	0.06	0.02	1.00
GirlsEnrollDiff	0.72	-0.09	0.07	0.00
PTRatioDiff	0.46	-0.04	0.05	0.00
TOTREVdiff	0.41	0.16	0.24	1.00
NSLPLunchDiff	0.13	-0.00	0.00	0.28
FTEdiff2	0.13	0.00	0.00	0.95

PIP = Posterior inclusion probability; Post Mean = Expected A Posteriori estimate; Cond.Pos.Sign = Probability that the sign of the estimate is positive conditional on inclusion in the model.

Marginal posterior density plots for boys

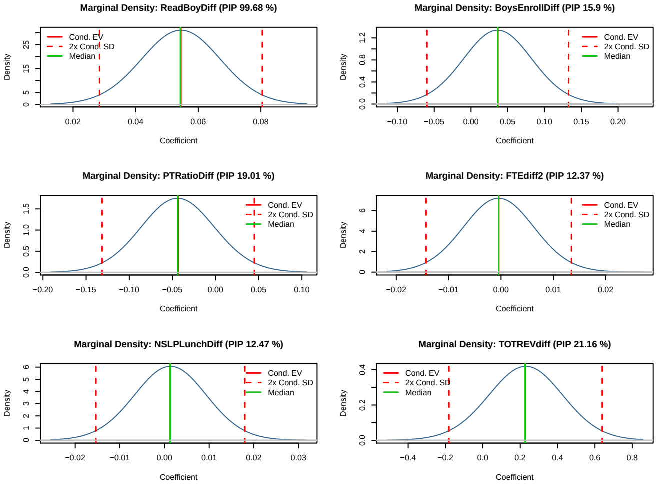


Figure 4: Marginal Posterior Density Plots and Posterior Inclusion Probabilities for Boys

Marginal posterior density plots for girls

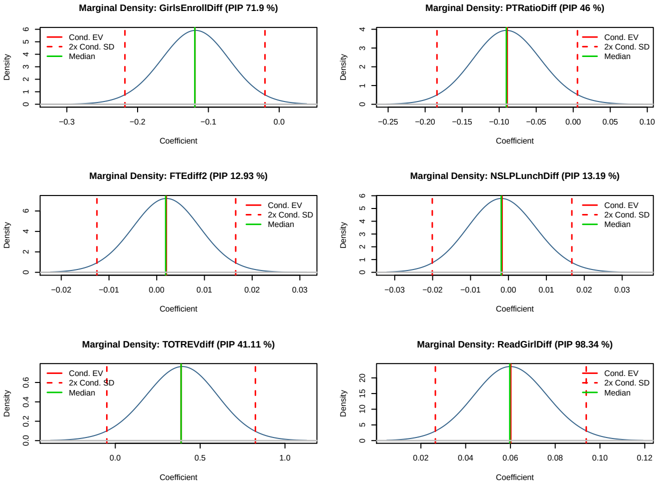
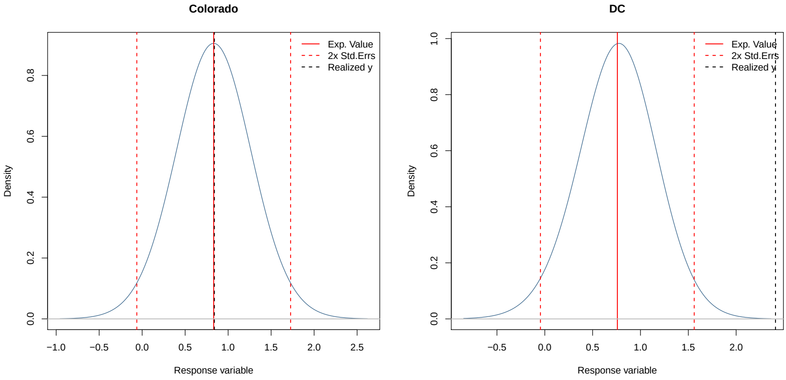


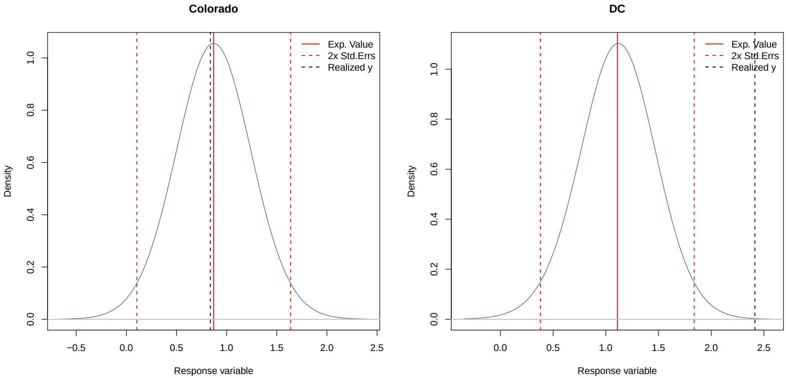
Figure 5: Marginal Posterior Density Plots and Posterior Inclusion Probabilities for Girls

Selected predictive densities

- Example of well-fitting (Colorado) and bad-fitting (DC) prediction models. Boys math achievement.



- Example of well-fitting (Colorado) and bad-fitting (DC) prediction models. Girls math achievement.



Summary

- We caution that the variables the data sets that were brought together to provide predictors of growth in 8^{th} grade mathematics achievement were not conceived or designed to provide policy relevant measures of growth over time.
- If there is a substantive interest in using state NAEP data for developing predictive models of growth in academic achievement outcomes, then it will be become necessary to consider the development of policy-relevant indicators specifically of growth.
- In this way, the full benefit of NAEP data can be leveraged for policy research and analysis.

Future directions

- Sensitivity to choice of parameter priors.
- Available priors are variants of Zellner's g-prior, including.
  - $g = N$ (Default unit information prior, used in this paper).
  - $g = \max(N, K^2)$ (Bayesian risk information prior).
  - $g = K^2$ (Risk information criterion).
  - $g = \log(N)$ (Hannan-Quinn prior).
  - g= EBL (Local empirical Bayes).

among others.

Future directions

- Sensitivity to choice of model priors.
  - $p(M_K) \propto 1$ (Default uniform prior).
  - Binomial model prior
  - Beta-binomial model prior

Future directions

- Evaluation of sensitivity based on scoring rules.
  - Log predictive score
  - Kullback-Leibler directed divergence
  - “Instrinsic discrepancy" (Bernardo & Reuda, 2002)
- $\mathcal{M}$-frameworks
- Applications to PISA (Kaplan & Jude, in progress)

Conclusions

- The question of using a model for some purpose beyond theory testing leads us to consider the accuracy of a model's predictions.
- BMA is known to yield models that perform better than any given sub-model on criteria of predictive accuracy.
- Paraphrasing G. E. P. Box's famous quote that "All models are wrong, but some are useful."
    - Not all models are equally good at prediction – yet all models contain some useful predictive information.
- This talk argues that BMA is a useful analytic approach to developing optimally predictive models in a wide variety of education research contexts where model uncertainty likely exists.

- Thanks to my collaborators
  - Dr. Cassie Chen (ETS)
  - Mingya Huang (PhD student, UW-Madison)
  - Dr. Danielle Lee (Liberty University)
  - Dr. Nina Jude (DIPF)
  - Sinan Yavuz (PhD student, UW-Madison)
- Portions of the research reported in this talk were supported by the Institute of Education Sciences, U.S. Department of Education (2011–2014), through Grant R305D110001 to the University of Wisconsin - Madison. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.
- Probabilistic forecasting work was conducted under a contract to the American Institutes for Research and the National Center for Education Statistics.
- Our published research papers, code, and data can be found at

http://bise.wceruw.org/index.html

THANK YOU